

# Heavy-traffic Asymptotics of Priority Polling System with Threshold Service Policy

Liu Zaiming, Chu Yuqing, Wu Jinbiao

*Department of Mathematics and Statistics, Central South University, Changsha, Hunan  
410083, PR China*

---

## Abstract

In this paper, by the singular-perturbation technique, we investigate the heavy-traffic behavior of a priority polling system consisting of three  $M/M/1$  queues with threshold policy. It turns out that the scaled queue-length of the critically loaded queue is exponentially distributed, independent of that of the stable queues. In addition, the queue lengths of stable queues possess the same distributions as a priority polling system with  $N$ -policy vacation. Based on this fact, we provide the exact tail asymptotics of the vacation polling system to approximate the tail distribution of the queue lengths of the stable queues, which shows that it has the same prefactors and decay rates as the classical  $M/M/1$  preemptive priority queues. Finally, a stochastic simulation is taken to test the results aforementioned.

*Keywords:*

Polling System, Heavy-traffic, Singular-perturbation, Tail Asymptotic, Stochastic Simulation

---

## 1. Introduction

The study of the two-queue priority polling system is motivated by its wide applications in computer and communication systems, such as ATM (Asynchronous Transfer Mode) switch systems and network standards like DQDB (Distributed Queue Dual Bus). ATM involves two different types of traffic: real time traffic (voice, video) and non-real time traffic (data), which

---

*Email addresses:* math\_lzm@csu.edu.cn (Liu Zaiming), chuyuqing@csu.edu.cn (Chu Yuqing), Corresponding author: wujinbiao@csu.edu.cn (Wu Jinbiao)

also need different types of QoS (Quality of Service) standard. By setting the threshold parameter, a higher priority is offered to real time traffic to shorten its delay and the delay of non-real time traffic is kept in a valid regime, which turns out to be a flexible way to control the operation of the whole system.

Lee and Sengupta first investigated the threshold-based priority systems in [1]. Later, a special case of two-queue  $M/M/1$  polling system with threshold policy was studied by Boxma, Koole and Mitrani in [2, 3]. The model was further extended with switch-over times by Deng et al. in [4, 5] and with one more server by Feng in [6].

In [7], we concerned with a three-queue model under threshold policy. The motivation stems from [8], in which Landry and Stavrakakis proposed a third type of traffic so called control traffic with Head-of-Line (HoL) in the integrated ATM environment, which involves critical network control and reservation information. In this paper, we focus on the heavy-traffic limits when there is a single critically loaded queue.

Using the singular-perturbation technique, we derive the lowest-order asymptotic of the joint queue-length distribution in terms of a small positive parameter measuring the closeness of the system to instability. The singular-perturbation technique was first applied to investigate the heavy-traffic behavior of interacting queues in [9]. Later, Boon and Winands [10] used this technique to a model with  $k$ -limited policies and presented the heavy-traffic behavior. It is noted that the singular-perturbation technique can be easily extended to a multi-queue system since it only needs the balance equations.

With the singular-perturbation technique, we conclude that the queue lengths in the stable queues have the same joint distribution as Model II, a preemptive priority polling system with  $N$ -policy vacation. In general, no closed-form expressions for the steady-state probabilities in Model II can be obtained. Using the Kernel method, which is reported detailedly in [11, 12], we present the exact tail asymptotics of queue lengths in Model II, which can further approximate the tail asymptotics of the stable queues.

The remainder of this paper is organized as follows. In Section 2, the model and some notations are introduced. In Section 3, the singular-perturbation technique is applied to derive the heavy-traffic limits and the detailed derivation is carried out in Section 4. In Section 5, we provide the exact tail asymptotics of queue lengths in Model II to approximate the tail asymptotics of the stable queues. In Section 6, a simulation is undertaken to evaluate the heavy-traffic asymptotics. We finally conclude the whole procedure and pro-

pose some topics for further research in Section 7.

## 2. Model Description

We consider a polling model with single server consisting of three queues  $Q_1, Q_2, Q_3$ . We refer to the customers queueing in  $Q_i$  as the type  $i$  customers,  $i = 1, 2, 3$ . The buffer capacity of each queue is infinite. Customers arrive at  $Q_i$  independently according to a Poisson process with rate  $\lambda_i$ . For type  $i$  customers, the service times are mutually independent and all follow an exponential distribution with rate  $\mu_i$ .  $Q_1$  has the HoL priority and  $Q_2$  has a higher priority over  $Q_3$ . In each queue customers are served according to FCFS discipline. We assume that the arrival processes and the service processes are independent. The service discipline is described as follows.

1.  $Q_1$  is served exhaustively, which means that the server serves the customers in  $Q_1$  until it is empty and then switches to  $Q_2$ ;
2. When the server is serving a customer in  $Q_2$ , if a type 1 customer arrives, then the server switches to  $Q_1$  immediately, otherwise, it continues serving the customers in  $Q_2$  until  $Q_2$  becomes empty and then switches to  $Q_3$ ;
3. When the server is serving a customer in  $Q_3$ , if a type 1 customer arrives, then the server switches to  $Q_1$  immediately, if the size of  $Q_2$  reaches a given threshold  $N$  and  $Q_1$  is empty, then the server switches to  $Q_2$  immediately, otherwise, it continues serving the customers in  $Q_3$  until  $Q_3$  becomes empty and then switches to  $Q_2$ .

It is assumed that all the switches are instantaneous. In addition, the switches caused by the threshold push the customer undergoing service to the head of the queue and the service of the interrupted customer resumes from the beginning.

The traffic load of  $Q_i$  is denoted by  $\rho_i = \lambda_i/\mu_i$ ,  $i = 1, 2, 3$ . We assume the ergodicity condition of the system  $\rho = \rho_1 + \rho_2 + \rho_3 < 1$  is satisfied.

Let  $X_i(t)$  be the number of customers in  $Q_i$  at time  $t$ , and  $S(t)$  be the position of the server at time  $t$  with  $S(t) \in \{1, 2, 3\}$ . The associated stochastic process  $\{Y(t), t \geq 0\} = \{(X_1(t), X_2(t), X_3(t), S(t)), t \geq 0\}$  is an aperiodic and irreducible four-dimensional Markov process. Let  $X_i$  ( $i = 1, 2, 3$ ) be the steady-state queue length of  $Q_i$  and  $S$  be the steady-state position of the server. Define the stationary probabilities:

$$p_s(x_1, x_2, x_3) = \lim_{t \rightarrow \infty} Pr\{Y(t) = (x_1, x_2, x_3, s)\}, \quad s = 1, 2, 3.$$

We study the heavy-traffic limits of the joint queue-length distribution by increasing the arrival rate  $\lambda_3$  so as to  $\rho \rightarrow 1^-$ , while keeping  $\lambda_1 \neq 0$ ,  $\lambda_2 \neq 0$  and  $\mu_1, \mu_2, \mu_3$  fixed. When  $\rho \rightarrow 1^-$ ,  $Q_3$  becomes critically loaded, whereas  $Q_1$  and  $Q_2$  remain stable since  $Q_1$  and  $Q_2$  have higher priorities over  $Q_3$ .

The single-perturbation technique is implemented here. We first apply a perturbation to  $\lambda_3$  in the balance equations, in which case  $Q_3$  is close to becoming critically loaded. Then we solve the lowest order terms in the balance equations to obtain the queue-length distributions of the stable queues  $Q_1$  and  $Q_2$ . At last we solve the first-order and second-order terms to get a differential equation and compute the scaled number of customers in  $Q_3$ .

Applying the Markov property, we obtain the following balance equations when  $x_3 \geq 2$ :

$$\begin{aligned}
& (\lambda_1 + \lambda_2 + \lambda_3 + \mu_1)p_1(x_1, x_2, x_3) \\
&= \lambda_1 p_1(x_1 - 1, x_2, x_3)\delta(x_1 \geq 2) + \lambda_2 p_1(x_1, x_2 - 1, x_3)\delta(x_2 \geq 1) \\
&\quad + \lambda_3 p_1(x_1, x_2, x_3 - 1) + \lambda_1 p_2(0, x_2, x_3)\delta(x_1 = 1, x_2 \geq 1) \\
&\quad + \mu_1 p_1(x_1 + 1, x_2, x_3) + \lambda_1 p_3(0, x_2, x_3)\delta(x_1 = 1, x_2 < N), \quad x_1 \geq 1, x_2 \geq 0,
\end{aligned} \tag{1}$$

$$\begin{aligned}
& (\lambda_1 + \lambda_2 + \lambda_3 + \mu_2)p_2(0, x_2, x_3) \\
&= \lambda_2 p_2(0, x_2 - 1, x_3)\delta(x_2 \geq 2) + \lambda_3 p_2(0, x_2, x_3 - 1) + \mu_1 p_1(1, x_2, x_3) \\
&\quad + \lambda_2 p_3(0, N - 1, x_3)\delta(x_2 = N) + \mu_2 p_2(0, x_2 + 1, x_3), \quad x_2 \geq 1,
\end{aligned} \tag{2}$$

$$\begin{aligned}
& (\lambda_1 + \lambda_2 + \lambda_3 + \mu_3)p_3(0, x_2, x_3) \\
&= \lambda_2 p_3(0, x_2 - 1, x_3)\delta(x_2 \geq 1) + \lambda_3 p_3(0, x_2, x_3 - 1) + \mu_3 p_3(0, x_2, x_3 + 1) \\
&\quad + [\mu_1 p_1(1, 0, x_3) + \mu_2 p_2(0, 1, x_3)]\delta(x_2 = 0), \quad 0 \leq x_2 \leq N - 1,
\end{aligned} \tag{3}$$

where  $\delta(\cdot)$  is Kronecker function.

In the above equations, we have omitted the parts for  $x_3 = 0$  and  $x_3 = 1$  which do not play a role after the perturbation since  $X_3$  tends to infinity as  $Q_3$  becomes critically loaded and the probability of  $Q_3$  being empty or 1 goes to zero.

Throughout the paper, we adopt the standard notations: a function  $F(x)$  is  $o(x)$  if  $F(x)/x \rightarrow 0$  as  $x \rightarrow 0$ ; a function  $F(x)$  is  $\mathcal{O}(x)$  if there exists a  $c \geq 0$  such that  $F(x)/x \rightarrow c$  as  $x \rightarrow 0$  while  $\mathcal{O}(1)$  is a constant time complexity; functions  $f(n)$  and  $g(n)$  of nonnegative integers  $n$ ,  $f(n) \sim g(n)$  means  $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 1$ .

### 3. Perturbation

From the stability condition the system becomes unstable as  $\rho_3 \rightarrow 1 - \rho_1 - \rho_2$ , i.e.  $\lambda_3 \rightarrow \mu_3(1 - \rho_1 - \rho_2)$ . Therefore it is assumed that

$$\lambda_3 = \mu_3(1 - \rho_1 - \rho_2) - \varepsilon\omega, \quad \omega > 0, 0 < \varepsilon \ll 1. \quad (4)$$

Let  $\zeta = \varepsilon x_3$ , and

$$p_s(x_1, x_2, x_3) = p_s(x_1, x_2, \zeta/\varepsilon) = \varepsilon \phi_{s,(x_1,x_2)}(\zeta, \varepsilon), \quad 0 < \zeta = \mathcal{O}(1), s = 1, 2, 3. \quad (5)$$

Taking (4) and (5) into the balance equations (1)-(3) and then taking the Taylor expansion, we obtain

$$\begin{aligned} & (\lambda_1 + \lambda_2 + \mu_1)\phi_{1,(x_1,x_2)}(\zeta, \varepsilon) \\ &= \lambda_1\phi_{1,(x_1-1,x_2)}(\zeta, \varepsilon)\delta(x_1 \geq 2) + \lambda_2\phi_{1,(x_1,x_2-1)}(\zeta, \varepsilon)\delta(x_2 \geq 1) \\ & \quad - (\mu_3(1 - \rho_1 - \rho_2) - \varepsilon\omega) \left( \varepsilon \frac{\partial \phi_{1,(x_1,x_2)}(\zeta, \varepsilon)}{\partial \zeta} - \frac{\varepsilon^2}{2} \frac{\partial^2 \phi_{1,(x_1,x_2)}(\zeta, \varepsilon)}{\partial \zeta^2} \right) \\ & \quad + \lambda_1\phi_{2,(0,x_2)}(\zeta, \varepsilon)\delta(x_1 = 1, x_2 \geq 1) + \mu_1\phi_{1,(x_1+1,x_2)}(\zeta, \varepsilon) \\ & \quad + \lambda_1\phi_{3,(0,x_2)}(\zeta, \varepsilon)\delta(x_1 = 1, x_2 < N) + o(\varepsilon^2), \quad x_1 \geq 1, x_2 \geq 0, \end{aligned} \quad (6)$$

$$\begin{aligned} & (\lambda_1 + \lambda_2 + \mu_2)\phi_{2,(0,x_2)}(\zeta, \varepsilon) \\ &= (\mu_3(1 - \rho_1 - \rho_2) - \varepsilon\omega) \left( -\varepsilon \frac{\partial \phi_{2,(0,x_2)}(\zeta, \varepsilon)}{\partial \zeta} + \frac{\varepsilon^2}{2} \frac{\partial^2 \phi_{2,(0,x_2)}(\zeta, \varepsilon)}{\partial \zeta^2} \right) \\ & \quad + \lambda_2\phi_{2,(0,x_2-1)}(\zeta, \varepsilon)\delta(x_2 \geq 2) + \mu_1\phi_{1,(1,x_2)}(\zeta, \varepsilon) \\ & \quad + \lambda_2\phi_{3,(0,N-1)}(\zeta, \varepsilon)\delta(x_2 = N) + \mu_2\phi_{2,(0,x_2+1)}(\zeta, \varepsilon) + o(\varepsilon^2), \quad x_2 \geq 1, \end{aligned} \quad (7)$$

$$\begin{aligned} & (\lambda_1 + \lambda_2)\phi_{3,(0,x_2)}(\zeta, \varepsilon) \\ &= (\mu_3(1 - \rho_1 - \rho_2) - \varepsilon\omega) \left( -\varepsilon \frac{\partial \phi_{3,(0,x_2)}(\zeta, \varepsilon)}{\partial \zeta} + \frac{\varepsilon^2}{2} \frac{\partial^2 \phi_{3,(0,x_2)}(\zeta, \varepsilon)}{\partial \zeta^2} \right) \\ & \quad + \mu_3 \left( \varepsilon \frac{\partial \phi_{3,(0,x_2)}(\zeta, \varepsilon)}{\partial \zeta} + \frac{\varepsilon^2}{2} \frac{\partial^2 \phi_{3,(0,x_2)}(\zeta, \varepsilon)}{\partial \zeta^2} \right) \\ & \quad + [\mu_1\phi_{1,(1,0)}(\zeta, \varepsilon) + \mu_2\phi_{2,(0,1)}(\zeta, \varepsilon)] \delta(x_2 = 0) \\ & \quad + \lambda_2\phi_{3,(0,x_2-1)}(\zeta, \varepsilon)\delta(x_2 \geq 1) + o(\varepsilon^2), \quad 0 \leq x_2 \leq N-1. \end{aligned} \quad (8)$$

It is noted that  $\lambda_3$  only plays a role in equations for  $\mathcal{O}(\varepsilon)$  terms and higher. Throughout the paper, we do Taylor expansions of  $\phi_{s,(x_1,x_2)}(\zeta, \varepsilon)$  ( $s = 1, 2, 3$ ) in powers of  $\varepsilon$  as follows

$$\phi_{s,(x_1,x_2)}(\zeta, \varepsilon) = \phi_{s,(x_1,x_2)}^{(0)}(\zeta) + \varepsilon \phi_{s,(x_1,x_2)}^{(1)}(\zeta) + o(\varepsilon^2), \quad s = 1, 2, 3. \quad (9)$$

In the next section the lowest order terms of the resulting equations after Taylor expansions are equated to find expressions for  $\phi_{s,(x_1,x_2)}^{(0)}(\zeta)$  ( $s = 1, 2, 3$ ), subsequently the first-order and second-order terms are equated to find the scaled queue-length distribution of  $Q_3$ .

For convenience, we introduce the corresponding probability generating functions(PGFs):

$$\begin{aligned}
Q_1^{(j)}(x, y, \zeta) &= \sum_{x_1=1}^{\infty} \sum_{x_2=0}^{\infty} \phi_{1,(x_1,x_2)}^{(j)}(\zeta) x^{x_1-1} y^{x_2}, \quad j = 0, 1, \\
Q_2^{(j)}(y, \zeta) &= \sum_{x_2=1}^{\infty} \phi_{2,(0,x_2)}^{(j)}(\zeta) y^{x_2-1}, \quad j = 0, 1, \\
Q_3^{(j)}(y, \zeta) &= \sum_{x_2=0}^{N-1} \phi_{3,(0,x_2)}^{(j)}(\zeta) y^{x_2}, \quad j = 0, 1, \\
Q_1(x, y, \zeta, \varepsilon) &= \sum_{x_1=1}^{\infty} \sum_{x_2=0}^{\infty} \phi_{1,(x_1,x_2)}(\zeta, \varepsilon) x^{x_1-1} y^{x_2}, \\
Q_2(y, \zeta, \varepsilon) &= \sum_{x_2=1}^{\infty} \phi_{2,(0,x_2)}(\zeta, \varepsilon) y^{x_2-1}, \\
Q_3(y, \zeta, \varepsilon) &= \sum_{x_2=0}^{N-1} \phi_{3,(0,x_2)}(\zeta, \varepsilon) y^{x_2}.
\end{aligned}$$

## 4. Model analysis

### 4.1. Equating the lowest-order terms

Equating the lowest-order terms of the resulting equation after the Taylor expansions of (6)-(8), we obtain

$$\begin{aligned}
&(\lambda_1 + \lambda_2 + \mu_1) \phi_{1,(x_1,x_2)}^{(0)}(\zeta) \\
&= \lambda_1 \phi_{1,(x_1-1,x_2)}^{(0)}(\zeta) \delta(x_1 \geq 2) + \lambda_2 \phi_{1,(x_1,x_2-1)}^{(0)}(\zeta) \delta(x_2 \geq 1) \\
&\quad + \lambda_1 \phi_{2,(0,x_2)}^{(0)}(\zeta) \delta(x_1 = 1, x_2 \geq 1) + \mu_1 \phi_{1,(x_1+1,x_2)}^{(0)}(\zeta) \\
&\quad + \lambda_1 \phi_{3,(0,x_2)}^{(0)}(\zeta) \delta(x_1 = 1, x_2 < N), \quad x_1 \geq 1, x_2 \geq 0, \tag{10}
\end{aligned}$$

$$\begin{aligned}
& (\lambda_1 + \lambda_2 + \mu_2)\phi_{2,(0,x_2)}^{(0)}(\zeta) \\
& = \lambda_2\phi_{2,(0,x_2-1)}^{(0)}(\zeta)\delta(x_2 \geq 2) + \mu_1\phi_{1,(1,x_2)}^{(0)}(\zeta) \\
& \quad + \lambda_2\phi_{3,(0,N-1)}^{(0)}(\zeta)\delta(x_2 = N) + \mu_2\phi_{2,(0,x_2+1)}^{(0)}(\zeta), \quad x_2 \geq 1, \tag{11}
\end{aligned}$$

$$(\lambda_1 + \lambda_2)\phi_{3,(0,x_2)}^{(0)}(\zeta) = \lambda_2\phi_{3,(0,x_2-1)}^{(0)}(\zeta), \quad 1 \leq x_2 \leq N-1, \tag{12}$$

$$(\lambda_1 + \lambda_2)\phi_{3,(0,0)}^{(0)}(\zeta) = \mu_1\phi_{1,(1,0)}^{(0)}(\zeta) + \mu_2\phi_{2,(0,1)}^{(0)}(\zeta). \tag{13}$$

We introduce  $P_0(\zeta)$  and  $\pi_{s,(x_1,x_2)}^{(0)}$  such that

$$\begin{aligned}
\phi_{s,(x_1,x_2)}^{(0)}(\zeta) &= \pi_{s,(x_1,x_2)}^{(0)}P_0(\zeta), \quad s = 1, 2, 3, \\
\sum_{x_1=1}^{\infty} \sum_{x_2=0}^{\infty} \pi_{1,(x_1,x_2)}^{(0)} + \sum_{x_2=1}^{\infty} \pi_{2,(0,x_2)}^{(0)} + \sum_{x_2=0}^{N-1} \pi_{3,(0,x_2)}^{(0)} &= 1.
\end{aligned}$$

Define

$$\begin{aligned}
L_1^{(0)}(x, y) &= \sum_{x_1=1}^{\infty} \sum_{x_2=0}^{\infty} \pi_{1,(x_1,x_2)}^{(0)} x^{x_1-1} y^{x_2}, \quad L_2^{(0)}(y) = \sum_{x_2=1}^{\infty} \pi_{2,(0,x_2)}^{(0)} y^{x_2-1}, \\
L_3^{(0)}(y) &= \sum_{x_2=0}^{N-1} \pi_{3,(0,x_2)}^{(0)} y^{x_2}.
\end{aligned}$$

Then it is clear that

$$Q_1^{(0)}(x, y, \zeta) = L_1^{(0)}(x, y)P_0(\zeta), \tag{14}$$

$$Q_2^{(0)}(y, \zeta) = L_2^{(0)}(y)P_0(\zeta), \tag{15}$$

$$Q_3^{(0)}(y, \zeta) = L_3^{(0)}(y)P_0(\zeta). \tag{16}$$

From (12), we get

$$L_3^{(0)}(y) = \sum_{x_2=0}^{N-1} (r_2 y)^{x_2} \pi_{3,(0,0)}^{(0)} = H(y) \pi_{3,(0,0)}^{(0)} = \beta(y) L_3^{(0)}(1), \tag{17}$$

where  $r_2 = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ ,  $H(y) = \frac{(r_2 y)^{N-1}}{r_2 y - 1}$  and  $\beta(y) = \frac{H(y)}{H(1)}$ .

Using the PGFs to rewrite the balance equations (10) and (11) leads to

$$xK(x, y)L_1^{(0)}(x, y) = \lambda_1 x[yL_2^{(0)}(y) + L_3^{(0)}(y)] - \mu_1 L_1^{(0)}(0, y), \tag{18}$$

$$ya(y)L_2^{(0)}(y) = \mu_1 L_1^{(0)}(0, y) - [\lambda_1 + \lambda_2(1 - y)]L_3^{(0)}(y), \quad (19)$$

where

$$K(x, y) = \lambda_1(1 - x) + \lambda_2(1 - y) + \mu_1 \left(1 - \frac{1}{x}\right),$$

$$a(y) = \lambda_1 + \lambda_2(1 - y) + \mu_2 \left(1 - \frac{1}{y}\right).$$

Clearly, for every  $|y| \leq 1$ , the kernel  $xK(x, y)$  has a unique zero:  $x = \alpha(y)$ . Applying the Kernel method to (18) and (19), it is easy to get

$$L_1^{(0)}(x, y) = \frac{\lambda_1 \mu_2 [x - \alpha(y)](y - 1)}{xK(x, y)[ya(y) - \lambda_1 y \alpha(y)]} L_3^{(0)}(y), \quad (20)$$

$$L_2^{(0)}(y) = \frac{\lambda_1(\alpha(y) - 1) + \lambda_2(y - 1)}{ya(y) - \lambda_1 y \alpha(y)} L_3^{(0)}(y), \quad (21)$$

Letting  $y \rightarrow 1$  and then letting  $x \rightarrow 1$  in (20) and (21), with L'Hôpital's rule, we obtain  $L_1^{(0)}(1, 1) = \frac{\rho_1}{1 - \rho_1 - \rho_2} L_3^{(0)}(1)$  and  $L_2^{(0)}(1) = \frac{\rho_2}{1 - \rho_1 - \rho_2} L_3^{(0)}(1)$ . By the normalizing condition, it is easy to get  $L_3^{(0)}(1) = 1 - \rho_1 - \rho_2$ . Therefore, we have  $L_1^{(0)}(1, 1) = \rho_1$  and  $L_2^{(0)}(1) = \rho_2$ . Moreover,

$$L_3^{(0)}(y) = \beta(y)(1 - \rho_1 - \rho_2). \quad (22)$$

It is not hard to see that equations (20)-(22) actually state an  $M/M/1$  preemptive priority polling system with  $N$ -policy vacation, denoted as Model II for short, described as follows:

There are two classes of customers in the system, the high- and low-priority customers, arriving independently according to two Poisson processes with rates  $\lambda_1$  and  $\lambda_2$ , respectively. Each class of customer is served according to the FCFS discipline. The server takes a vacation once the system empties and goes back to work once the size of the low-priority customers reaches  $N$  or there is a high-priority customer's arrival. The high-priority customers have preemptive priorities over the low-priority customers just like in the classical two-queue preemptive priority queueing system. Both classes of customers require an exponential amount of service times and are served with service rates  $\mu_1$  and  $\mu_2$ , respectively. All service times are independent and also independent of the arrival processes.

We determine the unknown expression of  $P_0(\zeta)$  in the rest of this section.



#### 4.2. Equating the first-order terms

In this subsection, by equating the first-order terms of the resulting equations after the Taylor expansion of the perturbed balance equations (6)-(8), we present an equation in Proposition 1.

##### Proposition 1.

$$(1 - \rho_1 - \rho_2) \left[ Q_1^{(1)}(1, 1, \zeta) + Q_2^{(1)}(1, \zeta) + Q_3^{(1)}(1, \zeta) \right] - Q_3^{(1)}(1, \zeta) \\ = - \left[ \frac{\mu_3}{\mu_1} \rho_1 + \frac{\mu_3}{\mu_2} \rho_2 \right] P_0'(\zeta).$$

PROOF. Taking the PGF of the first-order terms of the resulting equations after the Taylor expansion of (6)-(8), we have

$$xK(x, y)Q_1^{(1)}(x, y, \zeta) = \lambda_1 xyQ_2^{(1)}(y, \zeta) - \mu_1 Q_1^{(1)}(0, y, \zeta) + \lambda_1 xQ_3^{(1)}(y, \zeta) \\ - \mu_3 x(1 - \rho_1 - \rho_2)L_1^{(0)}(x, y)P_0'(\zeta), \quad (23)$$

$$ya(y)Q_2^{(1)}(y, \zeta) = \lambda_2 y^N \phi_{3, (0, x_2-1)}^{(1)}(\zeta) - \mu_1 Q_1^{(1)}(0, 0, \zeta) - \mu_2 Q_2^{(1)}(0, \zeta) \\ + \mu_1 Q_1^{(1)}(0, y, \zeta) - \mu_3 y(1 - \rho_1 - \rho_2)L_2^{(0)}(y)P_0'(\zeta), \quad (24)$$

$$[\lambda_1 + \lambda_2(1 - y)]Q_3^{(1)}(y, \zeta) = -\lambda_2 y^N \phi_{3, (0, x_2-1)}^{(1)}(\zeta) + \mu_1 Q_1^{(1)}(0, 0, \zeta) \\ + \mu_2 Q_2^{(1)}(0, \zeta) + \mu_3(\rho_1 + \rho_2)L_3^{(0)}(y)P_0'(\zeta). \quad (25)$$

Applying the Kernel method to (23)-(25), after some elementary calculations, we get

$$Q_1^{(1)}(x, y, \zeta) = \frac{\lambda_1[x - \alpha(y)]}{xK(x, y)} [yQ_2^{(1)}(y, \zeta) + Q_3^{(1)}(y, \zeta)] \\ - \frac{\mu_3(1 - \rho_1 - \rho_2)}{xK(x, y)} [xL_1^{(0)}(x, y) - \alpha(y)L_1^{(0)}(\alpha(y), y)]P_0'(\zeta). \quad (26)$$

$$[ya(y) - \lambda_1 \alpha(y)y]Q_2^{(1)}(y, \zeta) + [\lambda_1(1 - \alpha(y)) + \lambda_2(1 - y)]Q_3^{(1)}(y, \zeta) \\ = \mu_3 \left\{ (\rho_1 + \rho_2)L_3^{(0)}(y) - (1 - \rho_1 - \rho_2) \left[ yL_2^{(0)}(y) + \alpha(y)L_1^{(0)}(\alpha(y), y) \right] \right\} P_0'(\zeta). \quad (27)$$

Letting  $y \rightarrow 1$  and then letting  $x \rightarrow 1$  in (26), with L'Hôpital's rule, we obtain

$$Q_1^{(1)}(1, 1, \zeta) = \frac{\rho_1}{1 - \rho_1} [Q_2^{(1)}(1, \zeta) + Q_3^{(1)}(1, \zeta)] - \frac{\mu_3}{\mu_1} \frac{\rho_1(1 - \rho_1 - \rho_2)^2}{(1 - \rho_1)^2} P_0'(\zeta). \quad (28)$$

Letting  $y \rightarrow 1$  in (27) and using L'Hôpital's rule leads to

$$\begin{aligned} & \frac{1}{1-\rho_1} \left[ (1-\rho_1-\rho_2)Q_2^{(1)}(1, \zeta) - \rho_2 Q_3^{(1)}(1, \zeta) \right] \\ &= \left\{ -\frac{\mu_3}{\mu_1} \left[ \frac{\rho_1 \rho_2}{1-\rho_1} + \frac{\rho_1 \rho_2 (1-\rho_1-\rho_2)}{(1-\rho_1)^2} \right] - \frac{\mu_3}{\mu_2} \rho_2 \right\} P_0'(\zeta). \end{aligned} \quad (29)$$

From (28) and (29), we have

$$\begin{aligned} & (1-\rho_1-\rho_2) \left[ Q_1^{(1)}(1, 1, \zeta) + Q_2^{(1)}(1, \zeta) + Q_3^{(1)}(1, \zeta) \right] - Q_3^{(1)}(1, \zeta) \\ &= \frac{1}{1-\rho_1} \left[ (1-\rho_1-\rho_2)Q_2^{(1)}(1, \zeta) - \rho_2 Q_3^{(1)}(1, \zeta) \right] - \frac{\mu_3 \rho_1 (1-\rho_1-\rho_2)^2}{\mu_1 (1-\rho_1)^2} P_0'(\zeta) \\ &= - \left[ \frac{\mu_3}{\mu_1} \rho_1 + \frac{\mu_3}{\mu_2} \rho_2 \right] P_0'(\zeta). \end{aligned} \quad \square$$

#### 4.3. Equating the second-order terms

In this subsection we consider the sum of all  $\mathcal{O}(\varepsilon^2)$  terms in equations (6)-(8) to determine  $P_0(\zeta)$ .

Taking the summation over all  $x_1$  and  $x_2$  of (6)-(8), we get

$$\begin{aligned} & \mu_1 \sum_{x_2=0}^{\infty} \phi_{1,(1,x_2)}(\zeta, \varepsilon) \\ &= \lambda_1 \sum_{x_2=1}^{\infty} \phi_{2,(0,x_2)}(\zeta, \varepsilon) + \lambda_1 \sum_{x_2=0}^{N-1} \phi_{3,(0,x_2)}(\zeta, \varepsilon) - \mu_3 (1-\rho_1-\rho_2) \varepsilon \frac{\partial Q_1(1, 1, \zeta, \varepsilon)}{\partial \zeta} \\ &+ \left[ \omega \frac{\partial Q_1(1, 1, \zeta, \varepsilon)}{\partial \zeta} + \frac{\mu_3 (1-\rho_1-\rho_2)}{2} \frac{\partial^2 Q_1(1, 1, \zeta, \varepsilon)}{\partial \zeta^2} \right] \varepsilon^2 + \mathcal{O}(\varepsilon^3), \end{aligned} \quad (30)$$

$$\begin{aligned} & \lambda_1 \sum_{x_2=1}^{\infty} \phi_{2,(0,x_2)}(\zeta, \varepsilon) + \mu_2 \phi_{2,(0,1)}(\zeta, \varepsilon) \\ &= \mu_1 \sum_{x_2=1}^{\infty} \phi_{1,(1,x_2)}(\zeta, \varepsilon) + \lambda_2 \phi_{3,(0,N-1)}(\zeta, \varepsilon) - \mu_3 (1-\rho_1-\rho_2) \varepsilon \frac{\partial Q_2(1, \zeta, \varepsilon)}{\partial \zeta} \\ &+ \left[ \omega \frac{\partial Q_2(1, \zeta, \varepsilon)}{\partial \zeta} + \frac{\mu_3 (1-\rho_1-\rho_2)}{2} \frac{\partial^2 Q_2(1, \zeta, \varepsilon)}{\partial \zeta^2} \right] \varepsilon^2 + \mathcal{O}(\varepsilon^3), \end{aligned} \quad (31)$$

$$\begin{aligned}
& \lambda_1 \sum_{x_2=0}^{N-1} \phi_{3,(0,x_2)}(\zeta, \varepsilon) + \lambda_2 \phi_{3,(0,N-1)}(\zeta, \varepsilon) \\
&= \mu_1 \phi_{1,(1,0)}(\zeta, \varepsilon) + \mu_2 \phi_{2,(0,1)}(\zeta, \varepsilon) - \mu_3(1 - \rho_1 - \rho_2) \frac{\partial Q_3(1, \zeta, \varepsilon)}{\partial \zeta} + \mu_3 \frac{\partial Q_3(1, \zeta, \varepsilon)}{\partial \zeta} \\
&+ \left[ \omega \frac{\partial Q_3(1, \zeta, \varepsilon)}{\partial \zeta} + \frac{\mu_3(2 - \rho_1 - \rho_2)}{2} \frac{\partial^2 Q_3(1, \zeta, \varepsilon)}{\partial \zeta^2} \right] \varepsilon^2 + \mathcal{O}(\varepsilon^3);
\end{aligned} \tag{32}$$

Summing over (30)-(32), we obtain

$$\begin{aligned}
0 = & \left[ -\mu_3(1 - \rho_1 - \rho_2) \left( \frac{\partial Q_1(1, 1, \zeta, \varepsilon)}{\partial \zeta} + \frac{\partial Q_2(1, \zeta, \varepsilon)}{\partial \zeta} + \frac{\partial Q_3(1, \zeta, \varepsilon)}{\partial \zeta} \right) \right. \\
& + \mu_3 \frac{\partial Q_3(1, \zeta, \varepsilon)}{\partial \zeta} \left. \right] \varepsilon + \left[ \frac{\mu_3(1 - \rho_1 - \rho_2)}{2} \left( \frac{\partial^2 Q_1(1, 1, \zeta, \varepsilon)}{\partial \zeta^2} \right. \right. \\
& + \frac{\partial^2 Q_2(1, \zeta, \varepsilon)}{\partial \zeta^2} + \frac{\partial^2 Q_3(1, \zeta, \varepsilon)}{\partial \zeta^2} \left. \left. + \omega \left( \frac{\partial Q_1(1, 1, \zeta, \varepsilon)}{\partial \zeta} \right. \right. \right. \\
& + \frac{\partial Q_2(1, \zeta, \varepsilon)}{\partial \zeta} + \frac{\partial Q_3(1, \zeta, \varepsilon)}{\partial \zeta} \left. \left. + \frac{\mu_3}{2} \frac{\partial^2 Q_3(1, \zeta, \varepsilon)}{\partial \zeta^2} \right) \right] \varepsilon^2 + \mathcal{O}(\varepsilon^3),
\end{aligned} \tag{33}$$

Now taking the Taylor expansion (9) of equation (33), we obtain

$$\begin{aligned}
0 = & \left[ \mu_3(1 - \rho_1 - \rho_2) P_0''(\zeta) + \omega P_0'(\zeta) + \mu_3 Q_3^{(1)}(1, \zeta) - \mu_3(1 - \rho_1 - \rho_2) \times \right. \\
& \left. \left( Q_1^{(1)}(1, 1, \zeta) + Q_2^{(1)}(1, \zeta) + Q_3^{(1)}(1, \zeta) \right) \right] \varepsilon^2 + \mathcal{O}(\varepsilon^3) \\
& = \left[ \mu_3 \left( 1 - \rho_1 - \rho_2 + \frac{\mu_3}{\mu_1} \rho_1 + \frac{\mu_3}{\mu_2} \rho_2 \right) P_0''(\zeta) + \omega P_0'(\zeta) \right] \varepsilon^2 + \mathcal{O}(\varepsilon^3).
\end{aligned} \tag{34}$$

In (34), the first equation follows from (14)-(16) and the second equation follows from Proposition 1.

From the above derivation procedure, we can conclude the following Proposition.

**Proposition 2.** *After taking the summation over all  $x_1$  and  $x_2$  of the Taylor series of all perturbed balance equations (6)-(8), the  $\mathcal{O}(1)$  and  $\mathcal{O}(\varepsilon)$  terms*

cancel and, moreover, equating the  $\mathcal{O}(\varepsilon^2)$  terms yields the following differential equation for  $P_0(\zeta)$ :

$$\omega P_0'(\zeta) = - \left[ (1 - \rho_1 - \rho_2) + \frac{\mu_3}{\mu_1} \rho_1 + \frac{\mu_3}{\mu_2} \rho_2 \right] \mu_3 P_0''(\zeta).$$

#### 4.4. The scaled number of customers in the critically loaded queue

Now we can finally present the density of the scaled number of customers in  $Q_3$ , i.e.  $P_0(\zeta)$ . It can be obtained by combining the differential equation in Proposition 2 with  $\int_0^\infty P_0(\zeta) d\zeta = 1$  that

$$P_0(\zeta) = \eta e^{-\eta\zeta},$$

with  $\frac{\omega}{\eta} = \left[ 1 - \rho_1 - \rho_2 + \frac{\mu_3}{\mu_1} \rho_1 + \frac{\mu_3}{\mu_2} \rho_2 \right] \mu_3$ .

As a special case, we may take  $\omega = \mu_3$ , which gives  $\zeta = (1 - \rho)X_3$ , then

$$\frac{1}{\eta} = 1 - \rho_1 - \rho_2 + \frac{\mu_3}{\mu_1} \rho_1 + \frac{\mu_3}{\mu_2} \rho_2.$$

By applying the multiclass distributional law of Bertsimas and Mourtzinou [13] it directly follows that the scaled waiting time at  $Q_3$  follows an exponential distribution with parameter  $\mu_3\eta$ .

#### 4.5. Main result

**Theorem 1.** For  $\lambda_3 = \mu_3(1 - \rho_1 - \rho_2) - \varepsilon\omega$ , we have

$$\lim_{\varepsilon \downarrow 0} \mathbb{P}\{X_1 \leq x_1, X_2 \leq x_2, \varepsilon X_3 \leq \zeta\} = \mathcal{L}(x_1, x_2)(1 - e^{-\eta\zeta}),$$

where  $\mathcal{L}(\cdot, \cdot)$  is the joint cumulative distribution function(cdf) of the queue lengths of a preemptive priority polling system with  $N$ -policy vacation described in subsection 4.1.

The main result stated in Theorem 1 can be interpreted as follows: in the heavy-traffic regime,

- R1. The queue lengths in the stable queues have the same distribution as that of a preemptive priority polling system with  $N$ -policy vacation.
- R2. The scaled number of customers in the critically loaded queue is exponentially distributed with parameter  $\eta$ .

R3. The queue lengths in the stable queues and the (scaled) number of customers in the critically loaded queue are independent.

For R1, since  $Q_3$  is critically loaded,  $Q_3$  would be visited during each cycle. From the perspective of  $Q_1$  and  $Q_2$ , the server goes on a vacation once the server goes to  $Q_3$  when  $Q_1$  and  $Q_2$  are empty, and goes back to work once a type 1 customer arrives or there are  $N$  type 2 customers queueing, which actually is an  $N$ -policy vacation.

For R2, we note that the total workload in the system equals the amount of workload in an M/G/1 queue with arrival rate  $\lambda_1 + \lambda_2 + \lambda_3$  and hyper-exponentially distributed service times, i.e. the service time is exponentially distributed with parameter  $\mu_i$  with probability  $\frac{\lambda_i}{\lambda_1 + \lambda_2 + \lambda_3}$ ,  $i = 1, 2, 3$ . Based on the heavy-traffic results for the M/G/1 queue (see [13]), the distribution of the scaled total workload converges to an exponential distribution with mean  $\rho\mathbb{E}[R]$ , where  $R$  is a residual service time and

$$\mathbb{E}[R] = \frac{\frac{1}{\mu_1}\rho_1 + \frac{1}{\mu_2}\rho_2 + \frac{1}{\mu_3}\rho_3}{\rho}.$$

In the heavy traffic, since almost all customers are located in  $Q_3$ , the total number of customers at this queue is also exponentially distributed with mean  $\mu_3 \left( \frac{1}{\mu_1}\rho_1 + \frac{1}{\mu_2}\rho_2 + \frac{1}{\mu_3}\rho_3 \right)$ . Since  $\lambda_3 \uparrow \mu_3(1 - \rho_1 - \rho_2)$ , the scaled number of customers in  $Q_3$  is exponentially distributed with parameter  $\eta$ .

Finally, R3 follows from the time-scale separation in the heavy traffic which implies that the dynamics of the stable queues evolve at a much faster time scale than the dynamics of the critically loaded queue. Since the amount of “memory” of the stable queues asymptotically vanish compared to that of the critically loaded queue, the queue lengths in the stable queues are independent of the (scaled) number of customers in the critically loaded queue in the limit.

**Remark 1.** From the above procedure, it is easy to see that, when there is a single critically loaded queue in the heavy traffic, the stable queues with threshold policies can always be transferred into a priority polling system with  $N$ -policy vacation.

## 5. Exact tail asymptotics in Model II

In Section 4, we have derived the PGFs of the queue-length distributions of the stable queues, which have the same distributions as Model II. As

known, no closed-form expressions for the steady-state queue-length probabilities can be obtained. In this section, we carry out a detailed analysis on the exact tail asymptotics for the stationary distributions in Model II, which provides us an approximation of the stable queues.

### 5.1. Preliminary

First we introduce some necessary notations. The marginal distributions for the high- and low-priority customers are denoted by  $\pi_i^{(h)}$  and  $\pi_j^{(l)}$ , respectively. When  $j > 0$ , we write  $\pi_j^{(l)} = \pi_{1,j}^{(l)} + \pi_{2,j}^{(l)}$ , where  $\pi_{s,j}^{(l)}$  is the marginal distribution of the low-priority customers when the server is visiting  $Q_s$ ,  $s = 1, 2$ . We denote the distribution of the total number of customers by  $\pi_n^{(T)}$ . Let  $\lambda = \lambda_1 + \lambda_2$  and  $\bar{\rho}_1 = \lambda/\mu_1$ . Without loss of generality, throughout this section we assume that  $\lambda_1 + \lambda_2 + \mu_1 + \mu_2 = 1$ . To completely derive the exact tail asymptotics, we first introduce the following notations:

$$\begin{aligned}
b_1 &= \frac{\lambda_2}{\lambda_2 + (\sqrt{\mu_1} - \sqrt{\lambda_1})^2}, & b_2 &= \frac{\lambda_2}{\lambda_2 + (\sqrt{\mu_1} + \sqrt{\lambda_1})^2}, \\
\Delta(y) &= (\lambda + \mu_1 - \lambda_2 y)^2 - 4\lambda_1 \mu_1 = \lambda_2^2 (1 - b_1 y)(1 - b_2 y)/b_1 b_2, \\
x_1(y) &= \frac{(\lambda + \mu_1 - \lambda_2 y) - \sqrt{\Delta(y)}}{2\lambda_1} = \alpha(y), \\
x_2(y) &= \frac{(\lambda + \mu_1 - \lambda_2 y) + \sqrt{\Delta(y)}}{2\lambda_1}, \\
xK(x, y) &= -\lambda_1 x^2 + (\lambda + \mu_1 - \lambda_2 y)x - \mu_1 = -\lambda_1 (x - x_1(y))(x - x_2(y)), \\
c_0 &= \frac{(\lambda + \mu_1) - \sqrt{(\lambda + \mu_1)^2 - 4\lambda_1 \mu_1}}{2\mu_1}, & c_1 &= \frac{\lambda_2 c_0}{\sqrt{(\lambda + \mu_1)^2 - 4\lambda_1 \mu_1}}, \\
x_1 &= x_1(0) = \frac{c_0}{\rho_1}, & x_2 &= x_2(0) = \frac{1}{c_0}, \\
F(y) &= \lambda_2 y^2 - (1 - 2\mu_1 + \mu_2)y + 2\mu_2, \\
T^*(y) &= F(y) + y\sqrt{\Delta(y)}, & T(y) &= F(y) - y\sqrt{\Delta(y)}, \\
\eta_1 &= \frac{(1 - 2\mu_1) + \sqrt{(1 - 2\mu_1)^2 + 4(\mu_1 - \mu_2)\lambda_2}}{2\mu_2}, \\
\eta_2 &= \frac{(1 - 2\mu_1) - \sqrt{(1 - 2\mu_1)^2 + 4(\mu_1 - \mu_2)\lambda_2}}{2\mu_2}, \\
T(y)T^*(y) &= 4\mu_2^2(1 - y)(1 - \eta_1 y)(1 - \eta_2 y),
\end{aligned}$$

$$a = \frac{1 - \rho_1 - \rho_2}{2\mu_2} \frac{\eta_1}{\eta_1 - \eta_2}, \quad b = \frac{1 - \rho_1 - \rho_2}{2\mu_2} \frac{\eta_2}{\eta_2 - \eta_1},$$

$$D = (\lambda + \mu_1 - 2\sqrt{\lambda_1\mu_1})(\mu_1 - \mu_2 - \sqrt{\lambda_1\mu_1}) + \lambda_2\mu_2.$$

### 5.2. The PGFs of the stationary queue-length distribution

Define the following PGFs of the stationary queue-length distributions:

$$\psi_j^{(0)}(x) = \sum_{i=1}^{\infty} \pi_{1,(i,j)}^{(0)} x^{i-1}, \quad j = 0, 1, 2, \dots,$$

$$L^{(l)}(y) = \sum_{n=0}^{\infty} \pi_n^{(l)} y^n, \quad L^{(T)}(y) = \sum_{n=0}^{\infty} \pi_n^{(T)} y^n.$$

Now we present some Propositions to give the exact expressions of the PGFs defined above.

#### Proposition 3.

$$L_1^{(0)}(x, 1) = \frac{\rho_1(1 - \rho_1)}{1 - \rho_1 x}, \quad (35)$$

$$L_1^{(0)}(1, y) = \frac{\mu_2 - \lambda_2 y}{\lambda_2} L_2^{(0)}(y) - L_3^{(0)}(y), \quad (36)$$

$$L_1^{(0)}(y, y) = \frac{\lambda y - \mu_2}{\mu_1(1 - \bar{\rho}_1 y)} L_2^{(0)}(y) + \frac{\lambda}{\mu_1(1 - \bar{\rho}_1 y)} L_3^{(0)}(y), \quad (37)$$

$$L_1^{(0)}(x, 0) = \frac{c_0}{1 - c_0 x} L_3^{(0)}(0), \quad (38)$$

where  $L_1^{(0)}(x, y)$  and  $L_2^{(0)}(y)$  are expressed in (20) and (21) respectively.

PROOF. Adding (19) to (18), we have

$$xK(x, y)L_1^{(0)}(x, y) = y[\lambda_1 x - a(y)]L_2^{(0)}(y) + [\lambda_1(x - 1) + \lambda_2(y - 1)]L_3^{(0)}(y). \quad (39)$$

Then, letting  $y \rightarrow 1$ ,  $x \rightarrow 1$ ,  $x \rightarrow y$  and  $y \rightarrow 0$ , respectively, we get equations (35)-(38).  $\square$

#### Proposition 4.

$$\psi_0^{(0)}(x) = \frac{c_0}{1 - c_0 x} L_3^{(0)}(0), \quad (40)$$

$$\psi_j^{(0)}(x) = \frac{a_j}{1 - c_0 x} + \frac{\lambda_2 c_0 x}{\lambda_1 (1 - c_0 x)} \frac{\psi_{j-1}^{(0)}(x) - \psi_{j-1}^{(0)}(x_1)}{x - x_1}, \quad j = 1, 2, \dots \quad (41)$$

$$\text{where } a_j = \frac{c_0}{\lambda_1} \left[ \lambda_2 \psi_{j-1}^{(0)}(x_1) + \lambda_1 \left( \pi_{2,(0,j)}^{(0)} + \pi_{3,(0,j)}^{(0)} \delta(j < N) \right) \right].$$

PROOF. Equation (40) is obvious since  $\psi_0^{(0)}(x) = L_1^{(0)}(x, 0)$ . Using the PGFs to rewrite balance equation (10), we obtain

$$\psi_j^{(0)}(x) = \frac{\lambda_2 x \psi_{j-1}^{(0)}(x) + \lambda_1 x \left( \pi_{2,(0,j)}^{(0)} + \pi_{3,(0,j)}^{(0)} \delta(j < N) \right) - \mu_1 \psi_j^{(0)}(0)}{-\lambda_1 (x - x_1)(x - x_2)}. \quad (42)$$

Note that  $x_1 < 1$  and  $\psi_j^{(0)}(x)$  is analytic inside the unit circle, which implies that  $x_1$  is also a zero of the numerator of the righthand side of (42). Therefore,

$$\lambda_2 x_1 \psi_{j-1}^{(0)}(x_1) + \lambda_1 x_1 \left( \pi_{2,(0,j)}^{(0)} + \pi_{3,(0,j)}^{(0)} \delta(j < N) \right) = \mu_1 \psi_j^{(0)}(0). \quad (43)$$

Taking (43) into the numerator of the right hand side of (42) yields

$$\begin{aligned} \psi_j^{(0)}(x) &= \frac{\left[ \lambda_2 \psi_{j-1}^{(0)}(x_1) + \lambda_1 \left( \pi_{2,(0,j)}^{(0)} + \pi_{3,(0,j)}^{(0)} \delta(j < N) \right) \right] (x - x_1)}{-\lambda_1 (x - x_1)(x - x_2)} \\ &\quad + \frac{\lambda_2 x \left( \psi_{j-1}^{(0)}(x) - \psi_{j-1}^{(0)}(x_1) \right)}{-\lambda_1 (x - x_1)(x - x_2)}. \end{aligned}$$

Since  $x_2 = \frac{1}{c_0}$ , (41) can be obtained by simplifying the above equation.  $\square$

**Proposition 5.**

$$L_2^{(0)}(y) = \left[ \frac{aT^*(y)}{1 - \eta_1 y} + \frac{bT^*(y)}{1 - \eta_2 y} \right] \iota(y) \beta(y),$$

$$\text{with } \iota(y) = \frac{\mu_1 - \lambda + \lambda_2 y - \sqrt{\Delta(y)}}{2\mu_2(y-1)}.$$



PROOF. Simplifying (21), we get

$$\begin{aligned}
L_2^{(0)}(y) &= \frac{\lambda_1(x_1(y) - 1) + \lambda_2(y - 1)}{ya(y) - \lambda_1 y x_1(y)} L_3^{(0)}(y) \\
&= (1 - \rho_1 - \rho_2) \frac{2T^*(y)\mu_2(1 - y)}{T(y)T^*(y)} \frac{\lambda_1(x_1(y) - 1) + \lambda_2(y - 1)}{\mu_2(y - 1)} \beta(y) \\
&= \frac{1 - \rho_1 - \rho_2}{2\mu_2} \frac{T^*(y)}{(1 - \eta_1 y)(1 - \eta_2 y)} \iota(y) \beta(y) \\
&= \left[ \frac{aT^*(y)}{1 - \eta_1 y} + \frac{bT^*(y)}{1 - \eta_2 y} \right] \iota(y) \beta(y). \quad \square
\end{aligned}$$

**Proposition 6.**

$$L^{(T)}(y) = \left[ a \frac{T^*(y)}{1 - \eta_1 y} + b \frac{T^*(y)}{1 - \eta_2 y} \right] \kappa(y) \beta(y),$$

$$\text{with } \kappa(y) = \frac{2\mu_2(1-y) - (1-\mu_1 y)T(y)}{2\mu_1\mu_2 y(1-y)(1-\bar{\rho}_1 y)}.$$

PROOF. By the definition of  $L^{(T)}(y)$ , we have

$$\begin{aligned}
L^{(T)}(y) &= yL_1^{(0)}(y, y) + yL_2^{(0)}(y) + L_3^{(0)}(y) \\
&= \frac{1}{1 - \bar{\rho}_1 y} L_3^{(0)}(y) + \frac{(\mu_1 - \mu_2)y}{\mu_1(1 - \bar{\rho}_1 y)} L_2^{(0)}(y) \\
&= \left[ a \frac{T^*(y)}{1 - \eta_1 y} + b \frac{T^*(y)}{1 - \eta_2 y} \right] \kappa(y) \beta(y),
\end{aligned}$$

where the second equation follows from the expression (37) and the last follows the same idea used in Proposition 5.  $\square$

### 5.3. Analysis of singularities and asymptotic expansions

Along the same idea used for the classical priority model in [11], asymptotics of the coefficients are obtained using the following Tauberian-like theorem, which is Corollary 2 given in [14]. For a function  $f(y)$  that is analytic at  $y = 0$ , we denote the coefficient of  $y^k$  in the Taylor expression of  $f(y)$  by  $C_k[f(y)]$ .

For the compactness, we omit all the proofs in this subsection, which can be referred to [11].

**Lemma 1** (Flayolet and Odlyzko). *Assume that  $f(z)$  is analytic in  $\Delta(\phi, \varepsilon) = \{z : |z| \leq 1 + \varepsilon, |\text{Arg}(z - 1)| \geq \phi \text{ for } \varepsilon > 0 \text{ and } 0 < \phi < \pi/2\}$  except at  $z = 1$  and*

$$f(z) \sim K(1 - z)^s \quad \text{as } z \rightarrow 1 \quad \text{in } \Delta(\phi, \varepsilon).$$

*Then as  $n \rightarrow \infty$ :*

1. *If  $s \notin \{0, 1, 2, \dots\}$ ,*

$$f_n \sim \frac{K}{\Gamma(-s)} n^{-s-1}.$$

2. *If  $s$  is a nonnegative integer, then*

$$f_n = o(n^{-s-1}).$$

The key goal is to locate the dominant singularity, which determines the decay and to characterize the nature of the dominant singularity, which determines the prefactor and the singularity coefficient.

Define

$$\begin{aligned} \tilde{\Delta}(\phi, \varepsilon, a) = \{z : |az| \leq 1 + \varepsilon, |\text{Arg}(az - 1)| \geq \phi \text{ for } 0 < a < 1, \\ \varepsilon > 0 \text{ and } 0 < \phi < \pi/2\} - \{1/a\}. \end{aligned}$$

**Lemma 2.** *For the non-unit zeros  $1/\eta_1$  and  $1/\eta_2$ , we have*

1. *Both  $1/\eta_1$  and  $1/\eta_2$  are real.*
2.  *$\eta_1 > 0$ .*
3.  *$\eta_1 > \eta_2$ , and  $\eta_1 < |\eta_2|$  implies  $\eta_2 < 0$ .*
4.  *$\eta_2 < 0$ ,  $\eta_2 = 0$  or  $\eta_2 > 0$  if and only if  $\mu_2 < \mu_1$ ,  $\mu_2 = \mu_1$  or  $\mu_2 > \mu_1$ , respectively.*
5.  *$\eta_2 \neq b_1$ , and either  $T * (1/\eta_2) = 0$  or  $|\eta_2| < b_1$ .*

**Lemma 3** (Key Lemma). *There are three cases for the dominant singularity of  $L_2^{(0)}(y)$ :*

1. *If  $D > 0$ , then  $1 < 1/\eta_1 < 1/b_1$  and  $1/\eta_1$  is a zero of  $T(y)$  (but not  $T^*(y)$ ), and therefore  $1/\eta_1$  is the dominant singularity of  $L_2^{(0)}(y)$ , which is a simple pole.*
2. *If  $D = 0$ , then  $1 < 1/\eta_1 = 1/b_1$  and  $1/\eta_1$  is a zero of  $T(y)$  and  $T^*(y)$ , and therefore  $1/b_1$  is the dominant singularity of  $L_2^{(0)}(y)$ , which is both a branch point and a simple pole.*

3. If  $D < 0$ , then  $1 < 1/\eta_1 < 1/b_1$  and  $1/\eta_1$  is a zero of  $T^*(y)$  (but not  $T(y)$ ), and therefore  $1/b_1$  is the dominant singularity of  $L_2^{(0)}(y)$ , which is a branch point.

**Proposition 7.** *If  $\eta$  satisfies: (i)  $\eta \neq 0$ ; (ii)  $\eta \neq b_1$ ; (iii)  $|\eta| < b_1$  or  $T^*(\eta) = 0$ , then for  $\eta = \eta_i$ ,  $i = 1, 2$ ,*

$$C_n \left[ \frac{T^*(y)}{1 - \eta y} \iota(y) \beta(y) \right] \sim b_1 \beta(1/b_1) \sigma(\eta) n^{-3/2} b_1^n,$$

with  $\sigma(\eta) = \frac{K(\eta)}{b_1 \sqrt{\pi}}$  and  $K(\eta) = \frac{\lambda_2 b_1 \sqrt{1 - b_2/b_1}}{2 \sqrt{b_1 b_2} (\eta - b_1)}$ .

**Proposition 8.** *If  $\bar{\rho}_1 \geq 1$  and  $\eta$  satisfy: (i)  $\eta \neq 0$ ; (ii)  $\eta \neq b_1$ ; (iii)  $|\eta| < b_1$  or  $T^*(\eta) = 0$ , then for  $\eta = \eta_i$ ,  $i = 1, 2$ ,*

$$C_n \left[ \frac{T^*(y)}{1 - \eta y} \kappa(y) \beta(y) \right] \sim \beta(1/b_1) \sigma_1(\eta) n^{-3/2} b_1^n,$$

with  $\sigma_1(\eta) = \frac{K_1(\eta)}{b_1 \sqrt{\pi}}$  and  $K_1(\eta) = \frac{\lambda_2 b_1 \sqrt{1 - b_2/b_1} [(1 - \mu_1/b_1)(F(1/b_1) + 1) - 2\mu_2(1 - 1/b_1)]}{4\mu_1\mu_2\sqrt{b_1 b_2}(1 - \eta/b_1)(1 - 1/b_1)(1 - \bar{\rho}_1/b_1)}$ .

#### 5.4. Main results of exact tail asymptotics

In this subsection, we provide a complete exact tail asymptotics of the stationary distributions (the joint and marginal queue lengths and the total number of customers) by using the Tauberian-like Theorem to the related generating functions.

**Theorem 2.** *The exact tail asymptotics in the marginal stationary distribution  $\pi_n^{(h)}$  of the high-priority queue is given by*

$$\pi_n^{(h)} \sim (1 - \rho_1) \rho_1^n.$$

*The decay rate in the marginal distribution for the high-priority queue is  $\rho_1$ .*

PROOF. It is a direct consequence of the Taylor expansion of (35).  $\square$

**Theorem 3.** *The exact tail asymptotics in the joint stationary distribution along the high-priority queue is characterized by: for a fixed number  $j \geq 0$  of low-priority customers,*

$$\pi_{1,(n,j)}^{(0)} \sim \beta(0)(1 - \rho_1 - \rho_2) \left( \frac{c_1^j}{j!} \right) n^j c_0^{n-j}.$$

PROOF. First, by the induction, we prove

$$\psi_j^{(0)}(x) \sim c_0 \beta(0) (1 - \rho_1 - \rho_2) \left( \frac{c_1}{c_0} \right)^j \frac{1}{(1 - c_0 x)^{j+1}}, \quad j \geq 0, \text{ as } c_0 x \rightarrow 1. \quad (44)$$

It is true for  $j = 0$  since  $\psi_0^{(0)}(x) = \frac{c_0}{1 - c_0 x} L_3^{(0)}(0)$ . Assume that (44) is true for  $j = k$ , we then show it is true for  $j = k + 1$ . Rewrite equation (41) as

$$\psi_{k+1}^{(0)}(x) = \frac{a_{k+1}}{1 - c_0 x} + \frac{\lambda_2 c_0 x}{\lambda_1 (1 - c_0 x)} \frac{\psi_k^{(0)}(x) - \psi_k^{(0)}(x_1)}{x - x_1},$$

where  $a_{k+1}$  is a constant. Note that  $\frac{\lambda_2}{\lambda_1} \frac{c_0}{1 - x_1 c_0} = \frac{c_0}{c_1}$ . Hence,

$$\lim_{c_0 x \rightarrow 1} \frac{\psi_{k+1}^{(0)}(x)}{(1 - c_0 x)^{-(k+2)}} = c_0 \beta(0) (1 - \rho_1 - \rho_2) \left( \frac{c_1}{c_0} \right)^{k+1},$$

which is equivalent to (44). Therefore, (44) is true for all  $j \geq 0$ .

Applying Lemma 1 to (44), we have

$$\frac{C_n[\psi_j^{(0)}(x)]}{c_0^n} \sim c_0 L_3^{(0)}(0) \left( \frac{c_1}{c_0} \right)^j \frac{n^{(j+1)-1}}{\Gamma(j+1)} = c_0 L_3^{(0)}(0) \left( \frac{c_1}{c_0} \right)^j \frac{n^j}{j!}, \quad j \geq 0,$$

that is

$$\pi_{1,(n+1,j)}^{(0)} \sim L_3^{(0)}(0) \left( \frac{c_1}{c_0} \right)^j n^j c_0^{n+1-j}, \quad j \geq 0,$$

which completes the proof.  $\square$

**Theorem 4.** *The exact tail asymptotics in the joint stationary distribution along the low-priority queue is characterized by: for a fixed number  $i \geq 0$  of high-priority customers,*

1. (Exact geometric decay) *In the region of  $D > 0$ ,*

$$\pi_{2,(i,n)}^{(0)} \sim C_{2,l,1} [u(\eta_1)]^i \eta_1^n.$$

2. (Geometric decay with prefactor  $n^{-1/2}$ ) *In the region of  $D = 0$ ,*

$$\pi_{2,(i,n)}^{(0)} \sim C_{2,l,2} (\sqrt{\rho_1})^i n^{-1/2} b_1^n.$$

3. (Geometric decay with prefactor  $n^{-3/2}$ ) In the region of  $D < 0$ ,

$$\pi_{2,(i,n)}^{(0)} \sim C_{2,l,2}(1 + i\tilde{B})(\sqrt{\rho_1})^i n^{-3/2} b_1^n.$$

Here  $C_{2,l,1}$ ,  $C_{2,l,2}$ ,  $C_{2,l,3}$ ,  $u(\eta)$  and  $\tilde{B}$  are given below:

$$\begin{aligned} C_{2,l,1} &= 2aF(1/\eta_1)\beta(1/\eta_1), \\ C_{2,l,2} &= \frac{a\lambda_2\sqrt{1-b_2/b_1}}{\sqrt{\pi b_1}\sqrt{b_1 b_2}}\beta(1/b_1), \\ C_{2,l,3} &= [a\sigma(\eta_1) + b\sigma(\eta_2)]\beta(1/b_1), \\ u(\eta) &= \frac{1 - \mu_2 - (\lambda_2/\eta) - \sqrt{[1 - \mu_2 - (\lambda_2/\eta)]^2 - 4\lambda_1\mu_1}}{2\mu_1}, \\ \tilde{B} &= \frac{\mu_2 - \mu_1 - \mu_2 b_1 + \sqrt{\lambda_1\mu_1}}{\sqrt{\lambda_1\mu_1}}. \end{aligned}$$

PROOF. In the case of  $i = 0$ ,

1. if  $D > 0$ , then  $T(\frac{1}{\eta_1}) = 0$ , and we can prove  $\iota(\frac{1}{\eta_1}) = \eta_1$ . Hence,

$$\begin{aligned} &\lim_{\eta_1 y \rightarrow 1} \left[ \frac{L_2^{(0)}(y)}{(1 - \eta_1 y)^{-1/2}} \right] \\ &= a \lim_{\eta_1 y \rightarrow 1} T^*(y)\iota(y)\beta(y) + b \lim_{\eta_1 y \rightarrow 1} \left[ \frac{(1 - \eta_1 y)T^*(y)}{1 - \eta_2 y} \iota(y)\beta(y) \right] \\ &= aT^*(1/\eta_1)\iota(1/\eta_1)\beta(1/\eta_1) = 2aF(1/\eta_1)\eta_1\beta(1/\eta_1). \end{aligned}$$

Clearly,  $L_2^{(0)}(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, \eta_1)$ . By Lemma 1, we obtain

$$\pi_{2,(0,n+1)}^{(0)} \sim C_{2,l,1}\eta_1^{n+1}.$$

2. if  $D = 0$ , then  $T(\frac{1}{b_1}) = T(\frac{1}{\eta_1}) = 0$ , hence,  $\iota(\frac{1}{\eta_1}) = b_1$  and

$$\begin{aligned} &\frac{T^*(y)}{1 - \eta_1 y} \iota(y)\beta(y) \\ &= \frac{F(y) - F(1/b_1)}{1 - b_1 y} \iota(y)\beta(y) + \frac{y\sqrt{\Delta(y)}}{1 - b_1 y} \iota(y)\beta(y) \\ &\sim \frac{\rho_2 F'(1/b_1)\sqrt{1-b_2/b_1}}{2(1-b_1)\sqrt{b_1 b_2}} \sqrt{1-b_1 y} + \frac{\lambda_2\sqrt{1-b_2/b_1}\beta(1/b_1)}{\sqrt{b_1 b_2}\sqrt{1-b_1 y}} \\ &\sim \frac{\lambda_2\sqrt{1-b_2/b_1}\beta(1/b_1)}{\sqrt{b_1 b_2}\sqrt{1-b_1 y}}. \end{aligned}$$

Since  $\frac{T^*(y)}{1-\eta_1 y} \iota(y) \beta(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, b_1)$ , applying Lemma 1, we get

$$C_n \left[ \frac{T^*(y)}{1-\eta_1 y} \iota(y) \beta(y) \right] \sim \frac{\lambda_2 \sqrt{1-b_2/b_1} \beta(1/b_1)}{\sqrt{b_1 b_2} \sqrt{\pi}} n^{-3/2} b_1^{n+1}.$$

While with Proposition 7, we have

$$C_n \left[ \frac{T^*(y)}{1-\eta_2 y} \iota(y) \beta(y) \right] \sim \beta(1/b_1) \sigma(\eta_2) n^{-1/2} b_1^{n+1}.$$

Combining the above two asymptotics gives

$$\pi_{2,(0,n+1)}^{(0)} \sim C_{2,l,2} n^{-1/2} b_1^{n+1}.$$

3. if  $D < 0$ , the conclusion is a direct consequence of Proposition 7.

In the case of  $i > 0$ , the theorem can be proved by induction on  $i$ .

1. if  $D > 0$ , for  $i = 1$ , the balance equation is

$$\mu_1 \frac{\pi_{1,(1,n)}^{(0)}}{\eta_1^n} = (\lambda_1 + \lambda_2 + \mu_2) \frac{\pi_{2,(0,n)}^{(0)}}{\eta_1^n} - \frac{\lambda_2}{\eta_1} \frac{\pi_{2,(0,n-1)}^{(0)}}{\eta_1^{n-1}} - \mu_2 \eta_1 \frac{\pi_{2,(0,n+1)}^{(0)}}{\eta_1^{n+1}}.$$

It is easy to see that  $u(\eta)$  is the root of the equation with smaller module:  $\mu_1 [t(\eta)]^2 - [1 - \mu_2 - \lambda_2/\eta] t(\eta) + \lambda_1 = 0$ . Since  $T(\frac{1}{\eta}) = 0$ , we have  $u(\eta) = \frac{1-\mu_1-\mu_2\eta-\lambda_2/\eta}{\mu_1}$ . Therefore, we obtain

$$\pi_{1,(1,n)}^{(0)} \sim C_{2,l,1} A_1 \eta_1^n,$$

where  $A_1 = u(\eta_1)$ . Assume that for  $i \leq k$ ,

$$\pi_{1,(i,n)}^{(0)} \sim C_{2,l,1} A_i \eta_1^n.$$

Based on the balance equation

$$\begin{aligned} \mu_1 \pi_{1,(2,n)}^{(0)} &= (\lambda_1 + \lambda_2 + \mu_2) \pi_{1,(1,n)}^{(0)} - \lambda_2 \pi_{1,(1,n-1)}^{(0)} - \lambda_1 \pi_{2,(0,n)}^{(0)}, \\ \mu_1 \pi_{1,(k+1,n)}^{(0)} &= (\lambda_1 + \lambda_2 + \mu_2) \pi_{1,(k,n)}^{(0)} - \lambda_2 \pi_{1,(k,n-1)}^{(0)} - \lambda_1 \pi_{1,(k-1,n)}^{(0)}, \end{aligned}$$

and the inductive assumption  $\frac{\pi_{1,(k+1,n)}^{(0)}}{\eta_1^n} \rightarrow C_{2,l,1} A_{k+1}$ , we have

$$\mu_1 A_{k+1} = (\lambda_1 + \lambda_2 + \mu_2 - \frac{\lambda_2}{\eta_1}) A_k - \lambda_1 A_{k-1}, \quad k = 1, 2, 3, \dots$$

with  $A_0 = 1$  and  $A_1 = u(\eta_1)$ . Solving this difference equation leads to

$$A_k = [u(\eta_1)]^k, \quad k = 0, 1, 2, \dots$$

which gives the conclusion.

2. if  $D = 0$ , the proof is similar to that for case 1.
3. if  $D < 0$ , then  $u(b_1) = \sqrt{\rho_1}$ . Along the same idea in the proof of case 1, we get a difference equation

$$A_{k+1} = 2\sqrt{\rho_1}A_k - \rho_1 A_{k-1}, \quad k = 1, 2, 3, \dots$$

with  $A_0 = 1$  and  $A_1 = u(b_1)$ . Solving the equation yields the conclusion.  $\square$

**Theorem 5.** *The exact tail asymptotics in the marginal stationary distribution  $\pi_n^{(l)}$  of the low-priority queue is given by*

$$\pi_n^{(l)} = \frac{\mu_2}{\lambda_2} \pi_{2,(0,n+1)}.$$

PROOF. It is clear since  $L^{(l)}(y) = L_1^{(0)}(1, y) + yL_2^{(0)}(y) = \frac{\mu_2}{\lambda_2} L_2^{(0)}(y) - L_3^{(0)}(y)$ .  $\square$

**Theorem 6.** *The exact tail asymptotics in the stationary distribution  $\pi_n^T$  of total number of customers in the system is characterized below:*

*If  $\mu_1 = \mu_2$ , then*

$$\pi_n^T = \beta \left( \frac{1}{1 - \rho_1 - \rho_2} \right) (1 - \rho_1 - \rho_2)(\rho_1 + \rho_2)^n, \quad n = 0, 1, 2, \dots$$

*If  $\mu_1 \neq \mu_2$ , then*

1. *In the region of  $D > 0$ , three cases exist:*
  - a) *If (i)  $\bar{\rho}_1 \geq 1$ ; or (ii)  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 < \eta_1$ , then*

$$\pi_n^T \sim C_{t,1a} \eta_1^n.$$

- b) *If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 > \eta_1$ , then*

$$\pi_n^T \sim C_{t,1b} (\bar{\rho}_1)^n.$$

c) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 = \eta_1$ , then

$$\pi_n^T \sim C_{t,1c} n \eta_1^n.$$

2. In the region of  $D = 0$ , two cases exist:

a) If  $\bar{\rho}_1 \geq 1$ , then

$$\pi_n^T \sim C_{t,2a} n^{-1/2} b_1^n.$$

b) If  $\bar{\rho}_1 < 1$ , then

$$\pi_n^T \sim C_{t,2b} (\bar{\rho}_1)^n.$$

3. In the region of  $D < 0$ , three cases exist:

a) If  $\bar{\rho}_1 \geq 1$ , then

$$\pi_n^T \sim C_{t,3a} n^{-3/2} b_1^n.$$

b) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 \neq \sqrt{\rho_1}$ , then

$$\pi_n^T \sim C_{t,3b} (\bar{\rho}_1)^n.$$

c) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 = \sqrt{\rho_1}$ , then  $\bar{\rho}_1 = b_1 \neq \eta_1$  and

$$\pi_n^T \sim C_{t,3c} (\bar{\rho}_1)^n.$$

Here  $C_{t,1a}$ ,  $C_{t,1b}$ ,  $C_{t,1c}$ ,  $C_{t,2a}$ ,  $C_{t,2b}$ ,  $C_{t,3a}$ ,  $C_{t,3b}$  and  $C_{t,3c}$  are given below:

$$C_{t,1a} = \frac{(\mu_1 - \mu_2)\eta_1}{\mu_1(\eta_1 - \bar{\rho}_1)} C_{2,l,1},$$

$$C_{t,1b} = C_{t,2b} = C_{t,3b} = C_{t,3c} = \frac{(\mu_1 - \mu_2)}{\mu_1} \frac{1}{\bar{\rho}_1} L_2^{(0)}\left(\frac{1}{\bar{\rho}_1}\right) + L_3^{(0)}\left(\frac{1}{\bar{\rho}_1}\right),$$

$$C_{t,1c} = \frac{(\mu_1 - \mu_2)}{\mu_1} C_{2,l,1},$$

$$C_{t,2a} = \frac{\kappa(1/b_1)}{b_1} C_{2,l,2},$$

$$C_{t,3a} = [a\sigma_1(\eta_1) + b\sigma_1(\eta_2)]\beta(1/b_1).$$

PROOF. If  $\mu_1 = \mu_2$ , then  $L^{(T)}(y) = \frac{1}{1-\bar{\rho}_1 y} L_3^{(0)}(y)$  and  $\bar{\rho}_1 = \rho_1 + \rho_2$ . Hence, the conclusion is true. Now we consider the case  $\mu_1 \neq \mu_2$ .

1. In the region of  $D > 0$ ,



a) If (i)  $\bar{\rho}_1 \geq 1$ ; or (ii)  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 < \eta_1$ , then

$$\begin{aligned} \lim_{\eta_1 y \rightarrow 1} \left[ \frac{L^{(T)}(y)}{(1 - \eta_1 y)^{-1}} \right] &= a \lim_{\eta_1 y \rightarrow 1} \frac{(\mu_1 - \mu_2)y}{\mu_1(1 - \bar{\rho}_1 y)} T^*(y) \iota(y) \beta(y) \\ &= \frac{(\mu_1 - \mu_2)\eta_1}{\mu_1(\eta_1 - \bar{\rho}_1)} C_{2,l,1}. \end{aligned}$$

Since  $L^{(T)}(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, \eta_1)$ , applying Lemma 1, we get

$$\pi_n^{(T)} \sim C_{t,1a} \eta_1^n.$$

b) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 > \eta_1$ , then  $L^{(T)}(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, \bar{\rho}_1)$  and

$$\lim_{\bar{\rho}_1 y \rightarrow 1} \left[ \frac{L^{(T)}(y)}{(1 - \bar{\rho}_1 y)^{-1}} \right] = C_{t,1b}.$$

By Lemma 1, we have

$$\pi_n^{(T)} \sim C_{t,1b} (\bar{\rho}_1)^n.$$

c) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 = \eta_1$ , then  $L^{(T)}(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, \eta_1)$  and

$$\lim_{\eta_1 y \rightarrow 1} \left[ \frac{L^{(T)}(y)}{(1 - \eta_1 y)^{-2}} \right] = \frac{(\mu_1 - \mu_2)}{\mu_1} C_{2,l,1}.$$

By Lemma 1, we obtain

$$\pi_n^{(T)} \sim C_{t,1c} n \eta_1^n.$$

2. In the region of  $D = 0$ , two cases exist:

a) If  $\bar{\rho}_1 \geq 1$ , by Proposition 6,

$$L^{(T)}(y) = \left[ a \frac{T^*(y)}{1 - \eta_1 y} + b \frac{T^*(y)}{1 - \eta_2 y} \right] \kappa(y) \beta(y).$$

Similarly to the case of  $D = 0$  in Theorem 5, we have

$$\frac{T^*(y)}{1 - \eta_1 y} \kappa(y) \beta(y) \sim \frac{\lambda_2 \sqrt{1 - b_2/b_1} \kappa(1/b_1) \beta(1/b_1)}{\sqrt{b_1 b_2} \sqrt{1 - b_1 y}}.$$

In addition,  $\frac{T^*(y)}{1-\eta_1 y} \kappa(y) \beta(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, b_1)$ . Hence,

$$C_n \left[ \frac{T^*(y)}{1-\eta_1 y} \kappa(y) \beta(y) \right] \sim \frac{\lambda_2 \sqrt{1-b_2/b_1} \kappa(1/b_1) \beta(1/b_1)}{\sqrt{b_1 b_2} \sqrt{\pi}} n^{-1/2} b_1^n.$$

While with Proposition 8, we have

$$C_n \left[ \frac{T^*(y)}{1-\eta_2 y} \kappa(y) \beta(y) \right] \sim \beta(1/b_1) \sigma_1(\eta_2) n^{-3/2} b_1^{n+1}.$$

Combining the above two asymptotics leads to

$$\pi_n^T \sim C_{t,2a} n^{-1/2} b_1^n.$$

- b) If  $\bar{\rho}_1 < 1$ , then  $\bar{\rho}_1 > b_1$ . This can be proved by contradiction: if  $\bar{\rho}_1 = b_1$ , then  $\bar{\rho}_1 = \sqrt{\rho_1}$ , which follows from  $\bar{\rho}_1 - b_1 = \frac{-(\bar{\rho}_1 - \sqrt{\rho_1})^2}{\bar{\rho}_1 + 1 - 2\sqrt{\rho_1}}$ . After some manipulations, we get  $D = \mu_1(1 - \sqrt{\rho_1})^2(\mu_1 - \mu_2) \neq 0$ , which is contradict with  $D = 0$ . Hence,  $\bar{\rho}_1 > b_1$ . The remainder of the proof follows the same idea in the case 1-b).
- 3. In the region of  $D < 0$ , three cases exist:
  - a) If  $\bar{\rho}_1 \geq 1$ , then the conclusion follows from Proposition 8.
  - b) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 \neq \sqrt{\rho_1}$ , then  $\bar{\rho}_1 > b_1$ , the rest of the proof is similar to the case 1-b).
  - c) If  $\bar{\rho}_1 < 1$  and  $\bar{\rho}_1 = \sqrt{\rho_1}$ , then  $\bar{\rho}_1 = b_1 \neq \eta_1$ , we have

$$\lim_{\bar{\rho}_1 y \rightarrow 1} \left[ \frac{L^{(T)}(y)}{(1 - \bar{\rho}_1 y)^{-1}} \right] = L_3^{(0)}(1/\bar{\rho}_1) + \frac{\mu_1 - \mu_2}{\mu_1 \bar{\rho}_1} L_2^{(0)}(1/\bar{\rho}_1).$$

In addition,  $L^{(T)}(y)$  is analytic in  $\tilde{\Delta}(\phi, \varepsilon, \bar{\rho}_1)$ . By applying Lemma 1, we get

$$\pi_n^T \sim C_{t,3c}(\bar{\rho}_1)^n.$$

□

## 6. Stochastic simulation

This section tests our main results in Theorem 1 by comparing the ratio error of the waiting times and the cdfs of the queue lengths and waiting times. The ratio error was defined in [15] by

$$\text{Ratio error} = \frac{\text{Estimated value} - \text{Simulated value}}{\text{Simulated value}} \times 100\%,$$

Tab. 1 The ratio error of  $(1 - \rho)W_3$  for different loads

|       | $\rho = 0.8$ | $\rho = 0.9$ | $\rho = 0.95$ | $\rho = 0.975$ | $\rho = 0.99$ |
|-------|--------------|--------------|---------------|----------------|---------------|
| $E$   | -95.6000     | -91.1732     | -81.4272      | -72.7356       | -20.8699      |
| $Std$ | -93.4831     | -80.9886     | -68.5507      | -39.3165       | 32.0129       |

where the Estimated value is the result in Theorem 1 and the Simulated value is obtained by simulating under different traffic loads.

We consider a model with fixed parameters  $\lambda_1 = 0.1$ ,  $\lambda_2 = 0.3$ ,  $\mu_1 = 0.5$ ,  $\mu_2 = 1$ ,  $\mu_3 = 1.5$  and  $N = 10$ . We let  $\rho = 0.8, 0.9, 0.95, 0.975, 0.99$  to describe the procedure of  $\rho \rightarrow 1$  and  $\lambda_3$  can be determined by  $\lambda_3 = \mu_3(\rho - \rho_1 - \rho_2)$ . We use Matlab to undertake simulations under different traffic loads and each simulation runs until at least 10000 customers are served.

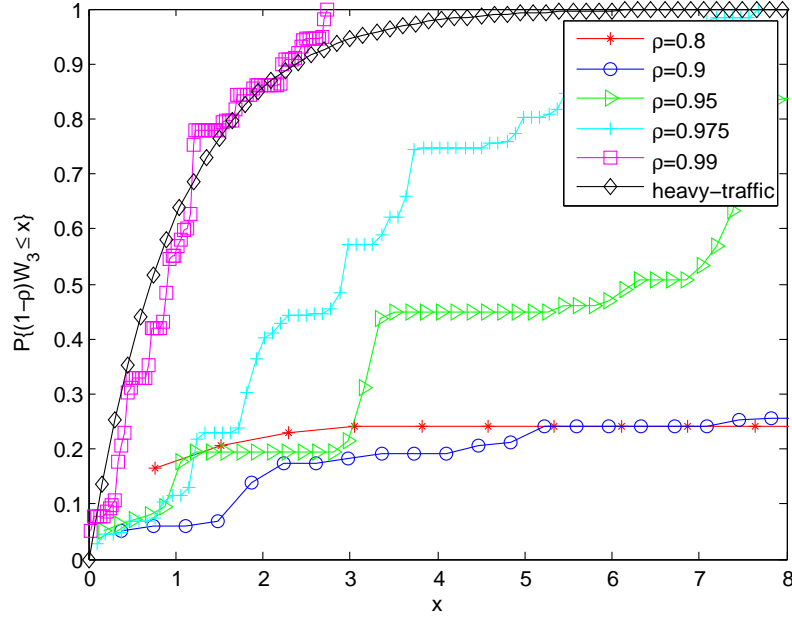


Fig. 1 The cdf of  $(1 - \rho)W_3$  for different loads

For this model, the scaled queue-length and waiting-time in the critically loaded queue are exponential distributed with parameter  $\eta$  and  $\mu_3\eta$  respectively in the heavy-traffic scenario. Fig.1 shows the cdf of  $(1 - \rho)W_3$  and

Tab.1 presents the ratio error of  $E(1 - \rho)W_3$  and  $Std(1 - \rho)W_3$ , where  $EX$  means the expectation of  $X$  and  $StdX$  means the standard deviation of  $X$ .

It is showed that the approximation performs well when  $\rho$  is very close to 1. However when  $\rho$  is moderate, the approximation seems not so accurate. This may own to the error of the simulation technique and the approximation theory since we only take the lowest order terms in the Taylor expansion. Fortunately, the higher-order terms can be obtained in the same procedure.

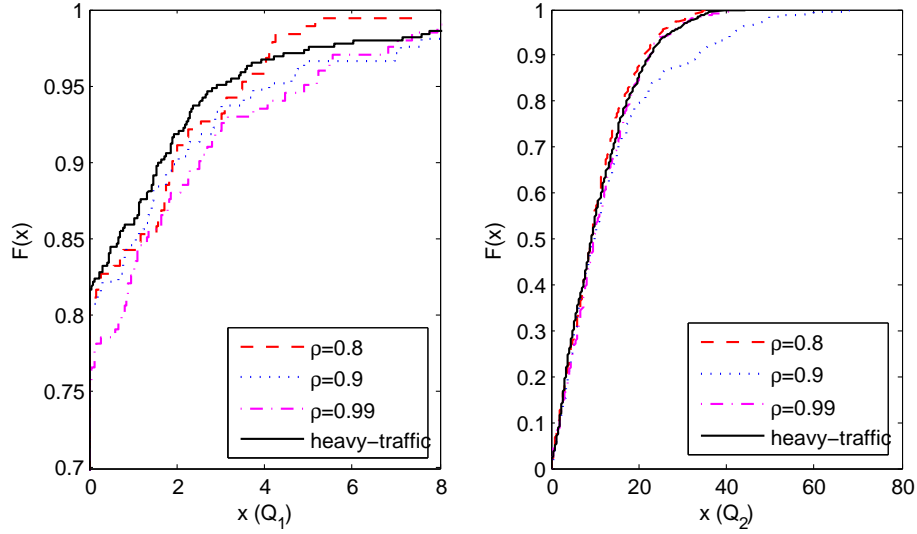


Fig. 2 Empirical cdf of waiting times in  $Q_1$  and  $Q_2$  for different loads

In the heavy-traffic regime, the queue lengths in the stable queues have the same distributions as that of a preemptive priority polling system with vacation, which is showed in Fig.2. From Fig.2, the distributions remain so closely whatever the traffic load  $\rho$  is, which can be explained by the preemptive priority service policy. The queue lengths in the stable queues are independent of the value of  $\rho$ , which may illustrate the conclusion that the queue lengths in the stable queues and the queue length in the critically loaded queue are independent. This can be showed more exactly in non-preemptive policy systems.

## 7. Conclusions

In this paper, we have derived the exact heavy-traffic limits of a three-queue priority polling system with threshold service policy using the singular-

perturbation technique. We also provided an approximation of the tail asymptotics of the stable queues, which describes the heavy-traffic behaviors more distinctly.

The singular-perturbation technique is based on the balance equation and hence can be extended to polling systems with multiple queues easily. It can be used to analyze the heavy-traffic limits of polling systems without multi-type branching properties [16]. However, if we apply the singular technique to the models with more than one critically loaded queues, then, initially, we may need to know the relative stabilities and, further, the degree of stability of each queue, which can be referred to [17]. In this way, we then find the most critically loaded queue and apply the technique. In addition, when  $\rho$  is moderate, the approximation seems not so accurate. Hence, it is necessary to seek for more efficient approximation techniques.

## References

- [1] D. S. Lee, B. Sengupta, Queueing analysis of a threshold based priority scheme for ATM networks, *IEEE/ACM Transactions on Networking (TON)* 1 (6) (1993) 709–717.
- [2] O. J. Boxma, G. M. Koole, I. Mitrani, A two-queue polling model with a threshold service policy, in: *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 1995. MASCOTS'95., Proceedings of the Third International Workshop on*, IEEE, 1995, pp. 84–88.
- [3] O. Boxma, G. Koole, I. Mitrani, Polling models with threshold switching, in: *Quantitative Methods in Parallel Systems*, Springer, 1995, pp. 129–140.
- [4] Y. Deng, J. Tan, Priority queueing model with changeover times and switching threshold, *Journal of Applied Probability* 38 (2001) 263–273.
- [5] Y. Deng, S. Song, J. Tan, Non-Preemptive Priority queueing model with changeover times and switching threshold, *Communication on Applied Mathematics and Computation* 15 (2001) 28–40.
- [6] F. Wei, M. Kowada, K. Adachi, Performance Analysis of a Two-queue Model with an (M, N)-threshold Service Schedule, *Journal of the Operations Research Society of Japan-Keiei Kagaku* 44 (2) (2001) 101–124.

- [7] Z. Liu, Y. Chu, J. Wu, On the Three-queue Priority Polling System with Threshold Service Policy, Submitted to Applied Mathematics and Computation.
- [8] R. Landry, I. Stavrakakis, Queueing study of a 3-priority policy with distinct service strategies, IEEE/ACM Transactions on Networking (TON) 1 (5) (1993) 576–589.
- [9] J. A. Morrison, S. C. Borst, Interacting queues in heavy traffic, Queueing Systems 65 (2) (2010) 135–156.
- [10] M. Boon, E. Winands, Heavy-traffic analysis of k-limited polling systems, Tech. rep., Technical Report 2013-002, Eurandom Preprint Series, 2013. To appear in Probability in the Engineering and Informational Sciences. Available at <http://www.eurandom.tue.nl/reports> (2013).
- [11] H. Li, Y. Q. Zhao, Exact tail asymptotics in a priority queue-Characterizations of the preemptive model, Queueing Systems 63 (1-4) (2009) 355–381.
- [12] H. Li, Y. Q. Zhao, Exact tail asymptotics in a priority queue-Characterizations of the non-preemptive model, Queueing Systems 68 (2) (2011) 165–192.
- [13] D. Bertsimas, G. Mourtzinou, Multiclass queueing systems in heavy traffic: An asymptotic approach based on distributional and conservation laws, Operations Research 45 (3) (1997) 470–487.
- [14] P. Flajolet, A. Odlyzko, Singularity analysis of generating functions, SIAM Journal on discrete mathematics 3 (2) (1990) 216–240.
- [15] T. L. Olsen, R. D. van der Mei, Polling systems with periodic server routing in heavy traffic: renewal arrivals, Operations Research Letters 33 (1) (2005) 17–25.
- [16] J. A. C. Resing, Polling systems and multitype branching processes, Queueing Systems 13 (4) (1993) 409–426.
- [17] L. Sum, R. K. Chang, Y. Xie, Relative stability analysis of multiple queues, in: Proceedings of the 1st international conference on Performance evaluation methodologies and tools, ACM, 2006, p. 65.